Designation: D 6582 – 00 (Reapproved 2005)

Standard Guide for Ranked Set Sampling: Efficient Estimation of a Mean Concentration in Environmental Sampling[1]

This standard is issued under the fixed designation D 6582; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This guide describes ranked set sampling, discusses its relative advantages over simple random sampling, and provides examples of potential applications in environmental sampling.

1.2 Ranked set sampling is useful and cost-effective when there is an auxiliary variable, which can be inexpensively measured relative to the primary variable, and when the auxiliary variable has correlation with the primary variable. The resultant estimation of the mean concentration is unbiased, more precise than simple random sampling, and more representative of the population under a wide variety of conditions.

1.3 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:* [2]

D 5792 Practice for Generation of Environmental Data Related to Waste Management Activities: Development of Data Quality Objectives

D 6044 Guide for Representative Sampling for Management of Waste and Contaminated Media

3. Terminology

3.1 *Definitions:*

3.1.1 *auxiliary variable*, *n*—the secondary characteristic or measurement of interest.

3.1.1.1 *Discussion*—In ranked set sampling, information contained in an auxiliary variable is useful for ranking the samples. This ranking may mimic the rankings of the samples with respect to the values of the primary variable when there is correlation between the auxiliary variable and the primary variable. Auxiliary information may include visual inspection, inexpensive quick measurement, knowledge of operational history, previous site data, or any other similar information.

3.1.2 *data quality objectives (DQO) process*, *n*—a quality management tool based on the scientific method and developed by the U.S. Environmental Protection Agency (EPA) to facilitate the planning of environmental data collection activities. **(D 5792)**

3.1.3 *equal allocation*, *n*—this occurs when the number of sets in ranked set sampling is an integer multiple of the size of the set.

3.1.4 *primary variable*, *n*—the primary characteristic or measurement of interest.

3.1.5 *ranked set sampling*, *n*—a sampling method in which samples are ranked by the use of auxiliary information on the samples and only a subset of the samples are selected for the measurement of the primary variable.

3.1.6 *representative sample*, *n*—a sample collected in such a manner that it reflects one or more characteristics of interest (as defined by the project objectives) of a population from which it is collected. **(D 6044)**

3.1.6.1 *Discussion*—A representative sample can be a single sample, a collection of samples, or one or more composite samples. A single sample can be representative only when the population is highly homogeneous. **(D 6044)**

4. Significance and Use

4.1 Ranked set sampling is cost-effective, unbiased, more precise and more representative of the population than simple random sampling under a variety of conditions **(1)**.[3]

4.2 Ranked set sampling (RSS) can be used when:

4.2.1 The population is likely to have stratification in concentrations of contaminant.

4.2.2 There is an auxiliary variable.

4.2.3 The auxiliary variable has strong correlation with the primary variable.

[1] This guide is under the jurisdiction of ASTM Committee D34 on Waste Management and is the direct responsibility of Subcommittee D34.01.01 on Planning for Sampling.

Current edition approved July 1, 2005. Published August 2005. Originally approved in 2000. Last previous edition approved in 2000 as D 6582-00.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[3] The boldface numbers in parentheses refer to the list of references at the end of this standard.

4.2.4 The auxiliary variable is either quick or inexpensive to measure, relative to the primary variable.

4.3 This guide provides a ranked set sampling method only under the rule of equal allocation. This guide is intended for those who manage, design, and implement sampling and analysis plans for management of wastes and contaminated media. This guide can be used in conjunction with the DQO process (see Practice D 5792).

5. Ranked Set Sampling (RSS)

5.1 Environmental sampling typically requires the identification of the locations where the samples are to collected. Subsequent analyses of these samples to quantify the characteristics of interest allow inference on the population mean concentration from the sample data.

5.2 A simple random sampling (SRS) approach is one sampling design that can be used. In this case, a set of random samples is identified and collected from a population and all of these samples are analyzed (for the primary variable).

5.3 Ranked set sampling (RSS) is similar to SRS in the identification and collection of the samples, but only a subset of the samples are selected for analysis. The selection is done by ranking the samples using auxiliary information on the samples and selecting a subset based on the rankings of the samples.

5.4 As can be seen from the steps described below, RSS is in fact a "stratified random sampling at the sample level," meaning that stratification of the population is induced after sampling and no construction of the strata is needed before sampling. Increased precision of stratified random sampling in the estimation of the population mean, relative to SRS, is well known, especially when the population is stratified by concentrations.

5.5 Increased precision of RSS relative to SRS means that the same precision can be achieved with fewer samples analyzed for the primary variable under RSS. RSS is therefore more cost-effective than SRS. When the objective is to minimize sampling and analytical costs, the number of samples can be determined so that RSS has precision equal to that of SRS at a lower cost.

5.6 The actual steps to conduct RSS are given below.

5.7 *Steps in Ranked Set Sampling (RSS)*:

5.7.1 Determine the total number of sample analyses (n) agreed to by the stakeholders. A planning process, such as a data quality objectives (DQO) process (Practice D 5792), may be used to determine this number.

5.7.2 Determine the primary variable and the auxiliary variable of interest.

5.7.3 Determine the size of the set, m. Study the auxiliary measurement and determine its capability in ranking the samples. For example, if the auxiliary measurement is visual inspection, its capability in ranking the samples may be somewhat limited. Namely, it may be capable of ranking 3–4 samples, but may have difficulty in ranking greater than 5 or 6 samples based on visual inspection; thus, the preferred size of the set (m) in ranked set sampling is about 3 or 4. On the other hand, an instrument-based quick-test may be capable of a larger m (see 5.14 for ranking criteria).

5.7.4 Calculate the needed number of replicates, r (the number of times the ranked sets are to be repeated). Divide n by m and round it up to whole number to obtain the needed r. Namely, $r = n/m$ and round up to whole number.

5.7.5 Randomly select a total of m^2r samples from the population, for example, by simple random sampling design, and randomly divide them into r replicates, with m^2 (m times m) samples in each replicate.

NOTE 1—In practice, the m^2r samples may not be taken all at once. More often, m random samples may be taken from a geographical sub-area of the population and are then ranked according to the auxiliary variable. This is repeated m times to obtain the first replicate of m^2 (m times m) samples. This entire process is repeated r times to obtain the needed r replicates.

5.7.6 Start with the first replicate of m^2 (m times m) samples. Arrange these samples into m sets of size m (an m by m matrix).

5.7.7 For each of the m sets in this replicate, rank the samples within each set by using the auxiliary measurement on the samples. When the observations on the auxiliary variable cannot be distinguished from each other, these observations are called "ties." Ties can be broken arbitrarily (namely, arbitrarily assigning one rank to one sample and a succeeding rank to the other).

5.7.8 Select samples for the measurement on the primary variable as follows. In set i, select and measure the sample with rank i, $i = 1, 2, ..., m$. Completion of this step leads to a total of m samples to be analyzed for the primary variable, out of a total of m^2 samples collected.

5.7.9 Repeat steps 5.7.6 through 5.7.8 for r times to obtain a total of $m \times r = n$ samples to be analyzed and measured for the primary variable.

5.8 Since the number of sets (m) in step 5.7.6 equals the size of the set (m), this is called equal allocation. RSS under unequal allocation tends to have additional gains in precision, relative to equal allocation; but, this gain is, in general, not large compared to the gain against SRS, and is not covered in this guide.

5.9 The value of n can be the total number of samples for which the budget can afford to analyze.

5.10 The rounding up in step 5.7.4 may cause the total number of analyses for the primary variable to exceed n. When this is the case, there are two options:

5.10.1 Obtain buy-in from the stakeholders to accept the slightly higher total number of sample analyses, or

5.10.2 Try different values of m and r to get the total number of analyses as close to n as possible.

5.11 *Estimation of Mean and Standard Error of the Mean*:

5.11.1 In 5.7, if $n = 12$, $m = 3$, and $r = 4$, the data on the primary variable obtained from the steps in that section may be summarized as in Table 1. The true mean concentration of the characteristic of interest is estimated by the arithmetic sample mean of the measured samples. For the hypothetical example in Table 1 (and assuming normal distribution of the data), the mean (M) is estimated as follows:

$$M = (X_{11} + X_{12} + X_{13} + X_{14} + X_{21} + X_{22} + ... + X_{34})/12 ..... \quad (1)$$

The standard error of the mean (S_M) is estimated as follows:

TABLE 1 Sample Values on Primary Variable

Set	Replicate	Value
1	1	X_{11}
	2	X_{12}
	3	X_{13}
	4	X_{14}
2	1	X_{21}
	2	X_{22}
	3	X_{23}
	4	X_{24}
3	1	X_{31}
	2	X_{32}
	3	X_{33}
	4	X_{34}

$$S_M = \{[(X_{11} - X_{1.})^2 + (X_{12} - X_{1.})^2 + (X_{13} - X_{1.})^2 + (X_{14} - X_{1.})^2 + (X_{21} - X_{2.})^2 + (X_{22} - X_{2.})^2 + (X_{23} - X_{2.})^2 + \ldots + (X_{34} - X_{3.})^2]/(m^2 r(r-1)\}^{1/2} \quad (2)$$

where:

X_{ij} = the value of the primary variable from the i^{th} set and the j^{th} replicate, and

X_i = the average of set i.

NOTE 2—The numerator of Eq 2 represents the squared differences between a value and its set average.

5.11.2 Given these estimates, inference about the population mean concentration can be made from the sample data (with some typical assumptions about the underlying statistical distribution of the data). This includes the use of confidence limits to estimate the population mean.

5.12 *An Illustration of RSS*:

5.12.1 An illustration of the steps in 5.7 and 5.11 is given in 5.12.2.1 through 5.12.2.9. The objective of this example is to estimate the mean of total petroleum hydrocarbon (TPH) concentration in the soil of a 1-acre site, down to the depth of one inch from the surface. Assume that the stakeholders agree that, due to cost and other considerations, a total of 12 analyses are the limit. Further assume that coloration on the surface of the soil is positively correlated with TPH concentration, with darker color indicating higher concentration.

5.12.2 The steps to carry out ranked set sampling are as follows:

5.12.2.1 $n = 12$, the desired total number of analyses for the primary variable.

5.12.2.2 The primary variable = TPH concentration and the auxiliary variable = soil coloration, where coloration is observed in-situ.

5.12.2.3 Professional judgment may indicate that visual inspection of soil coloration is capable of ranking three samples; thus, $m = 3$.

5.12.2.4 $r = n/m = 12/3 = 4$, the number of replicates.

5.12.2.5 Randomly select a total of m^2r samples = (3)(3)(4) = 36 samples. Divide these samples into 4 replicates with 9 samples in each replicate.

5.12.2.6 Take the first replicate of 9 samples and arrange them into 3 by 3 matrix; each row is called a set with set size of 3 (three samples in that set).

5.12.2.7 Rank the three samples within each of the three sets (namely, assigning ranks to the samples) according to the ranking of the soil coloration, giving a rank of 1 for lightest coloration, a rank of 3 for the darkest coloration.

5.12.2.8 In the first set, select the sample with rank 1 for the measurement of the primary variable. In set 2, select the sample with rank 2 for the measurement of the primary variable. And so forth for the third set. After this step, a total of $m = 3$ samples have been chosen for the measurement of the primary variable.

5.12.2.9 Repeat steps 5.12.2.6 through 5.12.2.8 for four times to obtain a total of $m \times r = 3 \times 4 = 12$ samples.

5.12.3 After steps 5.12.2.1 through 5.12.2.5, the 36 samples to be taken from the population may appear graphically as in Fig. 1. The samples in Fig. 1 are arbitrarily numbered from 1 through 36.

5.12.4 After steps 5.12.2.1 through 5.12.2.9, the rankings on the auxiliary variable and the measured values on the primary variable may appear as in Table 2. Each row of three samples can be called a cluster, and they are so designated in Table 2. These clusters can be marked in Fig. 1.

5.12.5 Table 3 is a summary of the data on the primary variable in Table 2. Note that the sample values of the primary variable and the bold-faced rank data in Table 2 happen to have the same ordering in all the replicates, except replicate 4, implying good correlation between the auxiliary variable and the primary variable.

5.12.6 When the data of the primary variable in Table 3 follow a normal distribution, the sample mean (M) and standard error of the mean (S_M) can be calculated as follows:

$$M = (9+10+12+15+15+16+20+14+17+18+23+20)/12 = 15.75, \quad (3)$$

$$S_M = \{[(9-11.5)^2 + (10-11.5)^2 + (12-11.5)^2 + (15-11.5)^2 + (15-16.25)^2 + (16-16.25)^2 + (20-16.25)^2 + (14-16.25)^2 + (17-19.5)^2 + (18-19.5)^2 + (23-19.5)^2 + (20-19.5)^2]/[(3^2)(4)(4-1)]\}^{1/2} = (62.75/81)^{1/2} = 0.76. \quad (4)$$

5.12.7 One- or two-sided confidence limits (CL) can be calculated from the sample mean and sample standard error of the mean (under the assumption that the data normally are distributed). For example, two-sided 95 % confidence limits are as follows:

$$CL = M \pm t_{0.95,11} S_M \quad (5)$$
$$= 15.75 \pm (2.201)\,(0.76)$$

	1	2	3	4	5	6	7	8	9	10	11
12	13	14	15	16	17	18	19	20			
21	22	23	24	25	26	27	28				
29	30	31	32	33	34	35	36				

FIG. 1 Thirty-Six Sample Locations in the Population

TABLE 2 Auxiliary Rankings[A] and Primary Variable of Thirty-Six Soil Samples in Four Replicates

Cluster	Replicate	Set	Sample Number[A]			Color Rank[A,B]			TPH
A	1	1	**1**	2	3	**1**	3	2	9
B	1	2	4	5	**6**	3	1	**2**	15
C	1	3	7	**8**	9	1	**3**	2	17
D	2	1	**10**	11	12	**1**	3	2	10
E	2	2	13	**14**	15	1	**2**	3	16
F	2	3	16	17	**18**	2	1	**3**	18
G	3	1	**19**	20	21	**1**	3	2	12
H	3	2	**22**	23	24	**2**	3	1	20
I	3	3	**25**	26	27	**3**	**1**	2	23
J	4	1	28	**29**	30	2	**1**	3	15
K	4	2	31	32	**33**	1	3	**2**	14
L	4	3	**34**	35	36	**3**	1	2	20

[A] Samples in bold-type are the ones to be analyzed for the primary variable.
[B] Samples are ranked within each set of three samples using soil coloration, where 1 = light, 2 = medium, and 3 = dark.

TABLE 3 Sample Values on Primary Variable (TPH, ppm)

Replicate	Sample Value for Set:		
	1	2	3
1	9	15	17
2	10	16	18
3	12	20	23
4	15	14	20
Set mean	11.50	16.25	19.50

$= 14.08$ *to* 17.42,

where the degrees of freedom for the *t*-statistic is (12-1) = 11.

NOTE 3—The standard error of the mean of the 12 sample values would have been 1.21 under SRS, much higher than the value of 0.76 under RSS. This improvement in precision in RSS is due to stratification "using an auxiliary variable."

5.12.7.1 In this illustration, the standard deviation under RSS is $(\sqrt{12})$ (0.76) = 2.64. If it is desired to determine the number of samples needed for RSS so that its precision (standard error) is equal to that of SRS, we can derive it by setting:

$$2.64/\sqrt{n} = 1.21 \tag{6}$$

Thus,

$$n = 5. \tag{7}$$

5.12.7.2 Thus, RSS can be designed so that it yields a total of five samples (for the primary variable). For example, a combination of $m = 3$ and $r = 2$ will yield a total of six samples. This will lead to the approximate level of desired precision. It can be seen that the six samples under RSS are only 50 % of the 12 samples needed under SRS for equal precision. The cost savings are obvious.

5.13 *Relative Advantages of Ranked Set Sampling*:

5.13.1 Ranked set sampling has the following advantages:

5.13.1.1 It provides an unbiased estimate of the population mean.

5.13.1.2 It is a more efficient estimator of the population mean than that from simple random sampling.

5.13.1.3 It has higher probability of obtaining representative samples relative to simple random sampling.

5.13.1.4 It exhibits superior performance in estimating the population mean for various underlying statistical distributions.

5.13.2 *Unbiased Estimate of the Mean*—According to Takahashi and Wakimoto (**2**), the ranked set sample mean is an unbiased estimate of the population mean. Dell and Clutter (**3**) show that it is an unbiased estimator regardless of imprecision in ranking using the auxiliary information.

5.13.3 *Relative Precision (RP) of RSS*:

5.13.3.1 The relative precision (RP) of RSS, relative to SRS, is the ratio of the variance of the SRS sample mean to the variance of the RSS sample mean. Namely,

$$\text{RP} = \frac{\text{Var(SRS sample mean)}}{\text{Var(RSS sample mean)}} \tag{8}$$

McIntyre (**4**) and Dell & Clutter (**3**) show that:

$$1 \le \text{RP} \le (m + 1)/2. \tag{9}$$

5.13.3.2 RP = 1 when ranking using auxiliary information has no value, meaning that the precision of RSS equals that of SRS. When ranking is near perfect, RP is very close to and not much below $(m + 1)/2$.

5.13.3.3 When RP is greater than 1, it means that the precision of RSS is better than that of SRS. Since the number of sets, *m*, in RSS is 2 or greater, RP is equal to 1.5 or greater. This means that the standard error of an RSS sample mean always is smaller than that of an SRS sample mean, given the same number of samples analyzed for the primary variable.

NOTE 4—When m=1, RSS is reduced to SRS and no auxiliary variable is needed.

5.13.3.4 This increased precision of RSS relative to SRS means that the same precision can be achieved with fewer samples (analyzed for the primary variable) under RSS. RSS is therefore more cost-effective than SRS under the stated conditions. For more details on the potential cost advantages of RSS over SRS, see 5.12.1 through 5.12.7.2 and Refs (**5 and 6**).

5.13.4 *Greater Representativeness of the Population*:

5.13.4.1 Lacayo and Neerchal (**7**) simulated elephant herds with different proportions of calves in the herd populations and find that RSS has smaller probabilities of obtaining nonrepresentative samples than SRS. This advantage should hold in environmental sampling, as well.

5.13.4.2 Sample representativeness of the population is an important environmental objective (Guide D 6044).

5.13.5 *Relative Precision Under Different Statistical Distributions*:

5.13.5.1 Patil, et al. (**1**) show that the variance of the sample mean under RSS is smaller than that obtained under SRS under a variety of statistical distributions for the population (including normal, lognormal, rectangular, exponential, gamma, Weibul, and triangular).

5.13.5.2 Some example relative precision (RP) results are given in Table 4.

5.14 *Ranking Criteria*:

5.14.1 Auxiliary variables that can be used to rank the samples within a set include the following:

5.14.1.1 Visual inspection.

5.14.1.2 Inexpensive quick measurement.

5.14.1.3 Knowledge of operational history.

TABLE 4 Relative Precision (RP) Under Different Population Distributions

Distribution		*m* in RSS	RP
Normal *N* (0, 1)		3	1.91
		5	2.77
Lognormal,	$\sigma = 0.1$	3	1.90
	$\sigma = 0.5$	3	1.70
	$\sigma = 1.0$	3	1.34
Lognormal,	$\sigma = 0.1$	10	4.73
	$\sigma = 0.5$	10	3.58
	$\sigma = 1.0$	10	2.06

where σ is the standard deviation for the underlying distribution.

5.14.1.4 Previous site data.

5.14.2 A quick method could be an in-situ measurement with an instrument. Operational history or previous site data may provide information about which areas are more highly contaminated than the others. In fact, stratified random sampling typically uses this kind of information to define the strata. When information on the stratification of the population is available in advance for the entire site, stratified random sampling may be more efficient than RSS, depending on how precise the knowledge of the population stratification is.

5.14.3 Near perfect ranking is achieved when the correlation between the auxiliary and primary variables approaches +1.0 or –1.0; however, perfect ranking is not needed for increased precision of RSS relative to SRS. Correlation close to +1.0 means that the ordering of the rank data on the auxiliary variable would be similar to the ordering of the data on the primary variable. Correlation close to –1.0 means that the ordering of the rank data on the auxiliary variable would be almost in reverse order of the primary variable. In this case of –1.0 correlation, the purpose of stratification "at the sample level" using the auxiliary variable also is achieved, thus leading to increased precision for RSS relative to SRS.

5.14.4 The relative efficiency (increase in precision) of RSS over SRS depends on ranking accuracy, the number of samples, and the underlying population distribution (**1**). Details on the last item are given in 5.13.2.

5.14.5 When there are ties and the ties are broken arbitrarily, this can reduce effectiveness of ranking, and thus, reduces the advantage of RSS to some extent. Tie-breaking, however, does not totally negate its advantage, nor does it introduce bias.

6. Uses of RSS in Environmental Sampling

6.1 RSS can be used when:

6.1.1 There is an auxiliary variable.

6.1.2 The auxiliary variable is either quick or inexpensive to measure, relative to the primary variable.

6.1.3 Cost of ranking the auxiliary variable is low relative to measurement of the primary variable.

6.1.4 The auxiliary variable has good correlation with the primary variable.

6.2 The advantages of RSS would be reduced or become negligible if any of the conditions listed in 6.1.1 through 6.1.4 are absent. For example, if the cost of measuring the auxiliary variable is large relative to that of measuring the primary variable, there may be an increase in total measurement costs. There would still be an increase in precision, however, relative to SRS (due to the effect of stratification using the auxiliary variable). In these cases when the advantages are not clear-cut, the stakeholder will need to weigh the pros and cons.

6.3 Nussbaum, et al. (**5**) describe the use of field RVP (Reid Vapor Pressure) as the auxiliary variable and the laboratory RVP as the primary variable in testing regulatory compliance of emissions at gas pumps, using RSS.

6.4 Pacific Northwest Laboratory (**8**) suggests the following potential applications:

6.4.1 Use of *in situ* readings (for example, readings of exterior drum conditions) to rank drums before a subset is selected for detailed inspection (on the assumption that the objective is to estimate the average condition of the drums).

6.4.2 In deciding which archived samples should be measured for a new contaminant, previous analytical results of these samples on other contaminants can be used to rank and select the samples for analysis of the new contaminant.

7. Keywords

7.1 auxiliary variable; bias; mean concentration; precision; primary variable; ranked set sampling; representativeness; sampling

REFERENCES

(1) Patil, G. P. et al. (1994), "Ranked Set Sampling," *In Handbook of Statistics 12*, G. P. Patil and N. P. Rao, eds, pp. 167–200, North-Holland, NY.

(2) Takahashi, K. and Wakimoto, K. (1968), "On Unbiased Estimates of the Population Mean Based on the Sample Stratified by Means of Ordering," Ann. of Institute of Stat. Math, 20:1–31.

(3) Dell, T. R. and Clutter, J. L. (1972), "Ranked Set Sampling Theory with Order Statistics Background," *Biometrics 28*, 545–555, June 1972.

(4) McIntyre, G. A. (1952), "A Method for Unbiased Selective Sampling, Using Ranked Sets," *Australian J. of Agricultural Research*, Vol 3, pp. 395–390.

(5) Nussbaum, B. D. et al (1997), "Cost Effective Gasoline Sampling Using Ranked Set Sampling," *1997 Proceedings of the Sec. on Statistics and the Environment*, American Statistical Association.

(6) Mode, N. A. et al. (1999), "Ranked Set Sampling for Ecological Research: Accounting for the Total Costs of Sampling," *Environmetrics*, Vol 10, pp. 179–194.

(7) Lacayo and Neerchal (1996), cited by Steven Millard and Nagaraj Neerchal in "Introduction to Environmental Statistics Using Environmental Stats for S-Plus," August 10, 1997, at *ASA Joint Statistical Meetings*, Anaheim, CA.

(8) Pacific Northwest Laboratory (1995), "DQO Statistics Bulletin," Vol 1, No. 1, June 1995, PNL-SA-26377.